CSE 415 Computer Science and Engineering, University of Washington

Name: Travis Xie

Document Classification with the Cosine Similarity

Consider the sentences A, B, and C below. First, stem any words having morphological inflections. Next eliminate stopwords in each. Finally determine the minimal reference vocabulary, sort it alphabetically, and encode each sentence as a vector of occurrence counts according to the reference vocabulary.

A. Musk says reserve your 30M ticket to Mars with a 1M deposit.
B. Muir loved the open space of the Mohave.
C. Mind, said Minsky, is a society.

Treat as stopwords: all prepositions, pronouns, articles, numbers or tokens beginning with numerals, and the verb "be".

deposit, love, Mars, Mind, Minsky, Mohave, Muir, Musk,
open, reserve, say, society, space, ticket

Suppose we are trying to decide whether A is more similar to B than to C. Compute the cosine similarity for A-B and for A-C. Which classification is implied by the results?

A o B: 0

A o C: 1

|| A || $\sqrt{6}$

|| B || $\sqrt{5}$

|| C || $\sqrt{4}$

cos(A, B): 0

cos(A, C): $= \frac{A \circ C}{\|A\| \|C\|} = \frac{1}{\sqrt{6} \cdot \sqrt{4}} \approx 0.204$

A: [1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1]

B: [0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0]

C: [0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0]